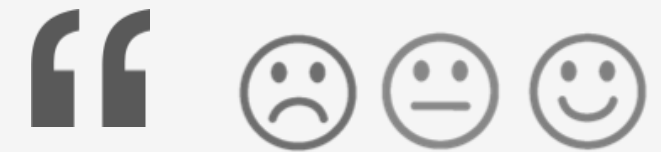


NLP 감정분석 기반
마케팅 시장 분석

키워드에 따른 커뮤니티 별 온라인 여론 감정 분석 시스템

빅데이터미네이터



커뮤니티별 이슈 키워드 감성 분석 시스템

각 커뮤니티의 제목, 본문, 댓글에 대한 크롤링을 진행하고, 이에 따른 주요 이슈 키워드를 추출한다.

또한 추출한 데이터를 바탕으로 감성 분석하여 결과를 도출하고자 한다.

커뮤니티별 익일 이슈 키워드와 그에 대한 감성 분석 결과를 사용자에게 알려주는 시스템이다.

1. Input

- 커뮤니티

: 네이트판, FM, 클리앙
(+DC, 뽀뿌)

- 크롤링 범위

→ 당일의 이슈 키워드 추출
(ex. 하루 중 일정 시간에 내용을
업데이트하여 사용자에게 알림)

: 제목, 본문, 댓글

2. 전처리

- 영어, 특수문자 등 제거
- 초성 (ㅋㅋ, ㅎㅎ 등) 제거

- 형태소 단위로 토큰화

→ 문장 단위로 순서는 유지해야함
(앞뒤 단어와의 연관성에 따른
감성 분석 가능)

- csv 파일로 저장 (프로토타입)

3. 프로토타입 구조

크롤링부 (+ 전처리)

↓
csv 파일 (추후 DB 구축)

↓ 감성어 사전
 (범용 사전 + 감성어 추가)

감성 분석기 => 결과 도출

가장 먼저 커뮤니티 선정한 커뮤니티인
네이트판, FM 코리아, 클리앙에 대한 크롤러를 만들었다.

제목, 본문, 댓글 내용에 대해 형태소 단위로 정리 한 뒤,
전체 형태소 개수, 명사 개수,
그리고 가장 많이 반복되는 키워드들을 뽑아보는 방식으로 진행했다.

커뮤니티 별 크롤링

01. 네이트판

- 당일 실시간 인기 톱 글이 올라오는 “툰커들의 선택 명예의 전당” 게시판의 제목, 본문, 댓글을 크롤링하였다.
- 제목 및 URL을 먼저 크롤링하고, 그 뒤 개별 본문으로 접근하여 각 본문 글과 댓글을 크롤링하는 순서로 진행했다.
- 크롤링한 데이터에 띄어쓰기가 올바르게 되지 않은 경우가 대다수여서 특수문자 등을 제거한 뒤, 한국어 전처리 패키지인 pykospacing을 이용하여 전체적으로 다시 띄어쓰기를 진행했다.
- KoNLPy 코엔엘파이, 한국어 형태소 분석기 내부의 Okt, Komoran, Twitter 등을 활용해 형태소 단위, 명사 단위의 분석을 진행하였다.

⇒ 앞으로 수정해 나가야하는 것

- 1) 전처리 단계에서 더 매끄러운 띄어쓰기 방식을 찾아볼 것
- 2) 형태소 분석기 사전에 온라인상 자주 사용되는 명사 등을 추가하여 전처리 단계에서 제외되지 않도록 해야 한다.
- 3) 더 의미 있는 최다 빈도 단어 추출을 위해 글자수 제한을 거는 등의 방법을 조사해 봐야한다.
- 4) 차단을 막기 위해 현재 sleep을 사용하고 있는데, 더 좋은 방법의 우회법을 찾아 봐야 한다.



전체 형태소 개수:
300613
전체 명사 개수:
35287

	단어	빈도수
0	남편	242
1	결혼	184
2	부모	156
3	엄마	153
4	여자	95
5	친구	81
6	남자	76
7	아이	74
8	본인	58
9	가요	56

```
#爬取
txt = []

for i in links:
    try:
        res = requests.get(i)
        res.raise_for_status()
        res.encoding = None
        html2 = res.text

        soup = BeautifulSoup(html2, 'html.parser')
        contentArea = soup.find("div", {'class': 'vlienarea'})
        paras = contentArea.findAll('div', {'id': 'contentArea'})

        content = ""

        for para in paras:
            content += para.text
            content = re.sub('[\r-\*\t-|a-zk-Z0-9]', '', content)
            content = re.sub('[=]', '', content)
            #content = re.sub('&nbsp;<br>|<br>', '', content)
            content = re.sub('[\W\d]', '', content)
            content = re.sub('[→★♥♡./!@#~+$.&*%->-|_!@#$%^&*~<->-]', '', content)
            content = spacing(content)

        txt.append(content)

except HTTPError as e:
    txt.append("")
except URLError as e:
    txt.append("")
except AttributeError as e:
    txt.append("")

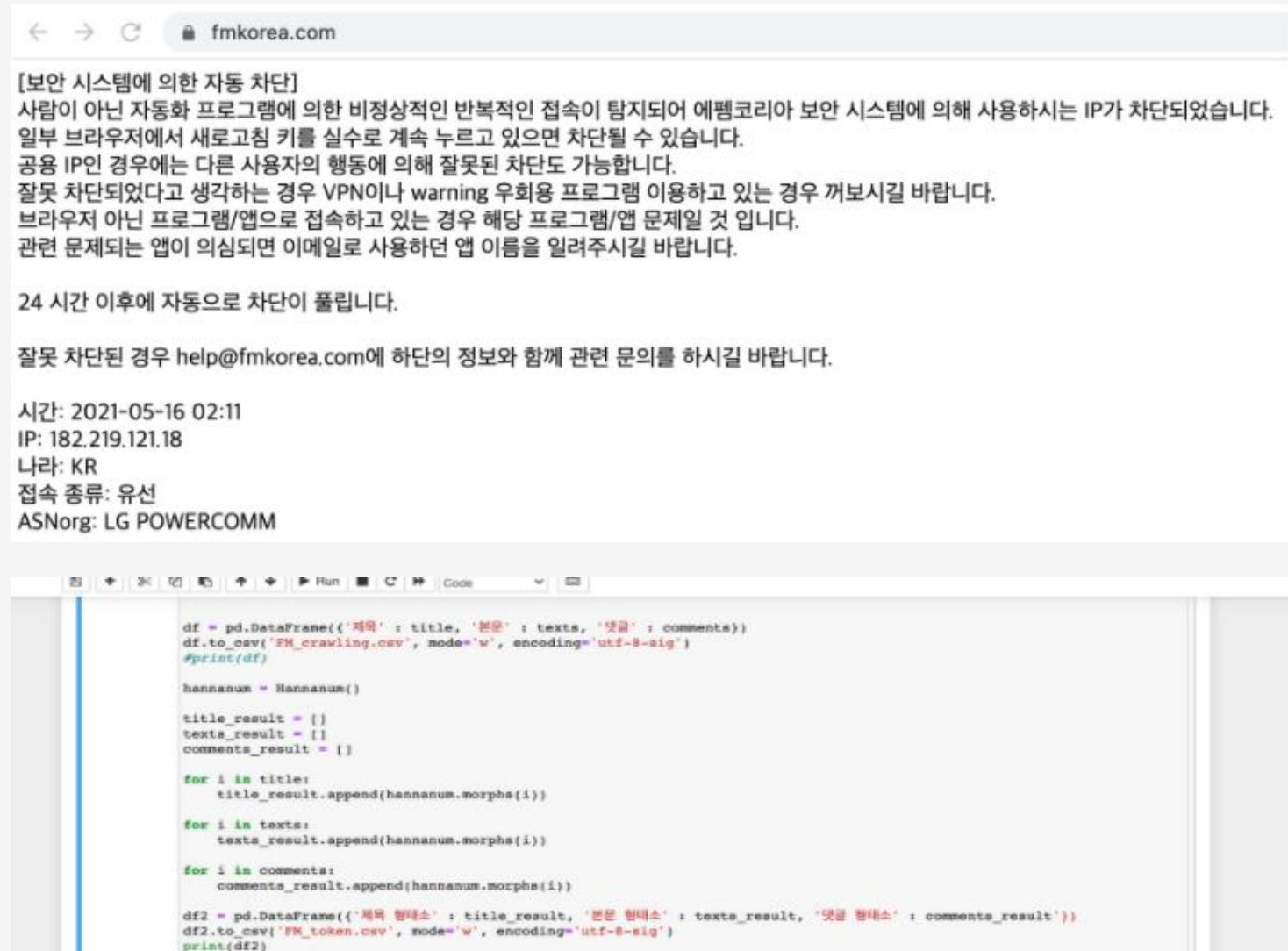
print(txt)
```

[illegible]

코드 일부

커뮤니티 크롤링

02. FM



- 본문 등 내용의 크롤링을 진행하다가, 차단되었다.
- 현재 다양한 우회 방법에 대해 자료조사를 하는 중이다.

이쯤되면 이 삼인방 재평가 시급한거 아닐????
 머니게임 한 번도 안 본 필름이 있으면 개추 ㅋㅋ;
 니가르 폭로때때 파이언크이 많아지자 코트 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
 머니게임..... 다시 보니 ㅋㅋ 선녀인 장면.gif
 머니게임) 3천 받아놓고 0원 3원 불러다가 지 세탁질 시켜놓고 말에 안드다고 뒤를수 맡겨?
 머니게임) 효감도 안가게 생긴 돌보장 의복인트 왜 넣은거냐?
 머니게임) 아니 파이가 3천 받았다고??
 머니게임 시즌2 최상의 멤버
 머니게임) 마지막 연영곡레딧 웃음질.jpg
 파이 퇴소이후 니가르 발언 해석.jpg
 머니게임)) 파이가 나간이유 뭣다...jpg
 머니게임 최종합을 보고 이해된 점
 오펜) 머니게임 최종 우승자
 20.05.15. 파이 오픈팅 토코 [머니게임 관련]
 이루리 조율히 있다가 6개월뒤에 커뮤 들면서 고소장쓸것같은
 니가르 폭로후 보니 소름돋는 장면
 실시간) 코트 채팅창
 배그하다 파이는 코트 들릴 영상 ㅋㅋㅋ
 n항할 3천미터 채팅본 코트방을 ㅋㅋㅋㅋㅋㅋ
 파이 : 코트오빠 원래 한 말에 책임 지시는 편인가요???
 ㅋㅋ파이가 개굴인거 같지?
 지을 니가르 폭로보고 제대로 느낀질.jpg
 머니게임 최종합헌 밝혀질
 머니게임 촬영 끝낸 당시... 파이...jpg
 실시간 파이 상황 ㅋㅋㅋ
 머니게임) 아까 니가르 방송에 이루리 남친 음.jpg
 머니게임 2주도 저렇게 힘든데
 씬 ㅋㅋㅋㅋ 머니게임 결말 예언한 놈 찾았다
 파이는 애는 진짜...
 니가르 폭로 8줄 요약 =====.jpg
 열팩좌 6년전 이루리 예언...
 머니게임)니가르:욕지말이 2천만원 요구함
 욕지말 유튜브 재개하자마자 다시 나락 위기
 니가르: 착한 사람 많지마세요. 이루리랑 대화 안한 이유
 머니게임중 시달 그만 올려라
 머니게임 실시간) 이루리 & N항에 대한 최종해명 정리
 머니게임)파이 더 역겨워진질 ㅋㅋ.jpg
 실시간 이루리 인스타 글쓴한 데려질 줄들ㅋㅋㅋㅋㅋㅋㅋㅋ .jpg
 니가르 안아주는 이루리 보고 쌍욕=====.avi
 질 몇천 번는 사람들이 거지 하고 유튜브를 놀락한거임 ㅋㅋ.jpg
 실시간 파이 맨날불과적전
 실시간.... 개굴 듀오
 다들주 펠코 폭발 =
 머니게임)) 돈 본래 오펜. TXT
 [머니게임] 근데 욕지말은 왜 욕하는거임?
 머니게임) 이루리... 진경성 보이던 장면.jpg
 니가르 가장 불쌍한 점
 어느회사 여자화장실 불라인드들
 머니게임)) 8화 결말 요약...jpg
 와 남들 파이 사과문 음 =====
 니가르: 카톡 내용까면 경찰서까지 가야한다
 니가르: ...

커뮤니티 크롤링

02. 클리앙



```
result_path = './result.xlsx'
result_word_path = './result_word.xlsx'

if __name__ == "__main__":
    try:
        make_excel(result_path)
        print('엑셀 파일 생성.')
    except PermissionError:
        print('엑셀 파일 생성 불가. 열려 있는 엑셀 파일을 종료해주세요.')
        sys.exit(1)
    print('=====')

    start_idx = 1
    with tqdm(total=100, ncols=80, desc='파일 생성 중') as pbar:
        for item in get_data():
            start_idx = append_excel(result_path, item, start_idx)
            pbar.update(1)

    print('=====')
    print('작업 완료.')
    input()

    try:
        make_excel2(result_word_path)
        print('엑셀 파일 생성.')
```

33	쫄지,마,항상,응원,합니다,,	
34	응원,합니다	
35	항상,응원,합니다	
36	응원,합니다	
37	응원,합니다	
38	응원,합니다	
39	2 장인,께서,영,면,하셨습니다,...	7년,전,아버지,소천,하셨을,때,도,클리앙,에서,많은,위로,를,받았
40	삼가,고인,의,명복,을,빕니다	
41	삼가,고인,의,명복,을,빕니다,/,Vollago	
42	삼가,고인,의,명복,을,빕니다	
43	삼가,고인,의,명복,을,빕니다	
44	평안,에,머우시길,빕니다	
45	삼가,고인,의,명복,을,빕니다	
46	글,에서,장인어른,에,대한,존경,과,사랑,이,느껴지네요,고인,의,명복,을,빕니다	
47	삼가고인의,명복,을,빕니다	
48	삼가,고인,의,명복,을,빕니다	
49	고인,이,남긴,삶,의,향기,를,느낄,수,있는,좋은,부고,글,인,것,같습니다,삼가,고인,의,명복,을,빕니다	
50	고인,의,명복,을,빕니다,좋은,곳,으로,가시길,...	
51	삼가,고인,의,명복,을,빕니다	
52	이제,평안,을,누리,시기를,기원,합니다	
53	삼가,고인,의,명복,을,빕니다	

분': 9, '재판': 3, '주위': 3, '생각': 3, '청년': 3, '지금': 3, '후': 2, '연락': 2,
2, '참고': 2, '감사': 2, '후원': 2, '저': 2, '스피커': 2, '양성': 2, '등': 2, '예
청소년': 2, '채널': 2, '준비': 2, '힘': 2, '어제': 1, '다스': 1, '출연': 1, '이야기
버': 1, '서도': 1, '신경': 1, '본의': 1, '분란': 1, '수': 1, '방송': 1, '응원': 1,
': 1, '비용': 1, '시민단체': 1, '사용': 1, '고양이': 1, '뉴스': 1, '개월': 1, '구
1, '무엇': 1, '문화': 1, '예술': 1, '정책': 1, '발굴': 1, '소통': 1, '가장': 1,

- 제목, 본문, 댓글을 크롤링하고 각각을 형태소 단위로 정리해보았다.

차주에는

1. 선정한 커뮤니티 3곳에 대한 크롤러 구현 진행
 2. 크롤러 구현 시 차단을 막기 위한 우회 방법 조사 및 추가
 3. 형태소 분석기 사전에 단어 추가하여 전처리 단계에서 제외되지 않도록 수정
 4. 감성 사전에 대한 학습 및 자료조사
- 을 진행하고자 합니다.



감사합니다 :))

빅데이터미네이터